Use of Multivariate Methods in the Analysis of Calculated Reaction Pathways

BJØRN K. ALSBERG, VIDAR R. JENSEN, and KNUT J. BØRVE*
Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway

ABSTRACT

It is suggested that multivariate methods such as principal component analysis (PCA) are useful tools in the analysis of large data sets from quantum chemical computations of reaction pathways. The potential of this methodology is investigated through an examination of the details of a medium-sized reaction: the Ziegler–Natta ethylene insertion reaction. Furthermore, PCA is used to compare two reaction pathways for the electrophilic addition of hydrochloric acid to propene. In both instances, the reaction pathways are determined at the Hartree–Fock level using the intrinsic reaction coordinate approach. The analyses are carried out on various kinds of descriptors, including geometry parameters, Mulliken charges, and overlap populations, and their relative efficiencies in terms of PCA modeling of the reactions are assessed. The results show that it may be necessary to combine analyses based on different descriptors and to analyze subsections of the reaction path separately in order to obtain both high resolution and interpretability. © 1996 by John Wiley & Sons, Inc.

Introduction

Over the past two decades, quantum chemical calculations have emerged as a useful and rather standard tool for obtaining chemical information. The steady improvement of computer technology has inspired the development of a series of new methods and programs applicable to chemical problems. Readily available program packages for quantum chemical calculations, like GAMESS-US,[1] GAUSSIAN,[2] and ADF,[3] are either free of charge or cheap if limited to academic use. In addition, knowledge about common methods for electronic structure calculations is increasing among nonspecialists. Thus, we are approaching a situation where general chemists frequently consider quantum chemistry as their method of choice when attacking a new problem. The systems treated by quantum chemistry are steadily growing in size as well as interest to chemists. Also increasing is the amount of information generated by the average quantum chemical calculation. Popular electronic structure programs now include code for calculating a number of properties, like electron density, electrostatic potential, and various kinds of population analyses. Except for the

*Author to whom correspondence should be addressed.

Journal of Computational Chemistry, Vol. 17, No. 10, 1197–1216 (1996)

CCC 0192-8651 / 96 / 101197-20

simplest problems, the amount of information generated can be very large. Observing complex covariance by simple inspection of the data matrices can be difficult and time consuming. Large portions of the data generated by quantum chemical calculations are probably not used at all, and there is thus a chance that important factors are left unnoticed.

An even more rapid increase in the amount of data has taken place within experimental chemistry, as the cost of extracting a large number of variables for each sample has decreased. The point in mind is exemplified by hyphenated spectroscopy and chromatography, as well as multidimensional spectroscopy. The need for extracting relevant information from multivariate measurements has led to the development of a wide variety of numerical methods for data processing and analysis; see, for example, Martens and Naes.[4] In some multivariate methods, the different kinds of covariance structures in large data sets are modeled by a few *latent variables*. These new variables are generally expressed as linear combinations of the original variables. The dimension of the problem is reduced, and inspection of the data is possible through low-dimensional plots that capture the essential information in the data. Examples of powerful multivariate methods are principal component analysis[5] (PCA) and partial-least-squares regression[6-8] (PLS).

It is thus tempting to apply multivariate techniques also in the analysis of results generated by quantum chemistry. So far, studies of this kind have focused on properties and conformations of static structures.[9-11] The information resulting from calculation of chemical reaction pathways constitutes an example of particularly large, theoretically generated data sets, and these trajectories should be suitable for treatment by multivariate techniques. PCA and other multivariate methods are already being used in the analysis of molecular dynamics trajectories. For instance, the collective motions of a protein molecule as simulated by Newtonian dynamics has been analyzed by projection of the calculated trajectories onto two-dimensional spaces spanned by the principal components.[12,13] These projections provide valuable insight about the regions (clusters) in conformational space that are occupied during the simulation. For systems studied by quantum chemistry, the projection technique may also be used in the analysis of properties extracted from the wave function along the reaction path. The main purpose of the present work is to explore the performance of multivariate methods, represented by PCA, in the analysis of geometrical parameters and properties from calculated reaction pathways.

Scheme for Analyzing Structural Data along a Reaction Path

In the present context, a reaction may be described by an ordered sequence of K nuclear *configurations* or *geometries* $\mathbf{G}^{(k)}, k = 1, 2, \ldots, K$, with corresponding numerical values $P_j^{(k)}, j = 1, 2, \ldots, J$, for J selected *descriptors* $\{P_j\}$. The term "descriptor" is used coherently with QSAR literature.[14-16] However, only descriptors which relate to geometrical structures or properties computed from the electronic wave function, will be considered.

The geometries are chosen to be evenly spaced along the reaction path, and as such they constitute a discrete representation of the pathway for the reaction studied. The reaction path itself is a special case of a *trajectory* on the potential energy surface of the system. Even though the path coordinate along the trajectory is 1-dimensional, it is *embedded* in a $(3N - 6)$-dimensional space $(3N - 5$ for linear systems), where N is the number of atoms in the system. In principle, the reaction pathway may be described by a smaller set of coordinates, given that the new coordinates are judiciously chosen. This idea is depicted in Figure 1 for a case with three internal coordinates ($N = 3$)

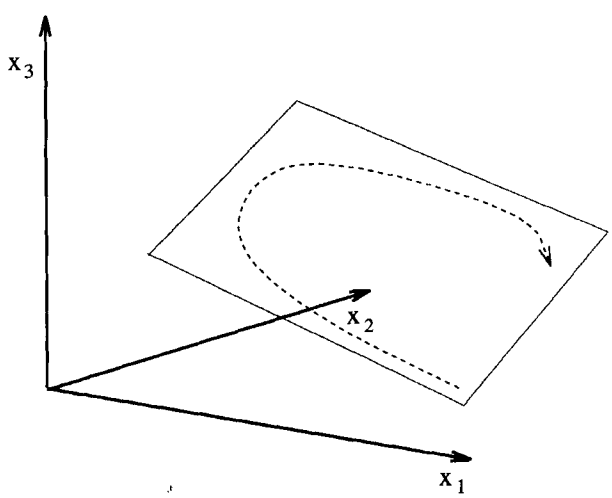

FIGURE 1. Illustration of the main idea in the present paper: A reaction path is embedded in a space spanned by three (e.g., internal) coordinates. In this example, a useful description of the variance along the pathway is obtained by projection onto a two-dimensional plane.

and a trajectory that can be described in terms of a two-dimensional embedding plane.

If one allows for very complex coordinates (relative to, say, a prechosen set of internal coordinates), then the reaction path length s would be the optimal choice of coordinate since it facilitates a one-dimensional plot of the energy profile of the reaction. However, for interpretative purposes, one would rather prefer a simple relationship between the new coordinates and those initially chosen. In this paper, it is suggested that useful low-dimensional embedding subspaces may be generated by means of new coordinates that are linearly related to the original coordinates, and with constant expansion coefficients. It is not to be expected that these subspaces give an exact embedding of the reaction path, but they may constitute very good *approximations* for interpretive purposes.

Hitherto the discussion has been kept in terms of structural parameters. However, these considerations are equally valid for the wider class of descriptors defined above.

An efficient and much used method to estimate embedding subspaces is *principal component analysis*.[5,17] In the PCA method, multivariate structures are represented in terms of a smaller set of new variables called *principal components* (PC). These new variables or components are oriented along the directions of maximum variance in the data set. As such, the principal components reveal structures that lie latent but concealed in the data, and the PCs are accordingly also termed *latent variables*. The details of the mathematical structure of the PCA algorithm is well known in the literature and will not be repeated here. Rather, we will try to establish an intuition for the basic quantities in PCA in the present context of analyzing a chemical reaction.

SCORES AND LOADINGS

Let $\mathbf{X}$ be the original data matrix of size $[K \times J]$. Row vector x_{k*} in this matrix consists of information about the reacting system at the kth geometry $\mathbf{G}^{(k)}$. Column vector x_{*j} shows how the value realized for descriptor P_j is changing along the reaction path.

The average vector $\bar{\mathbf{x}} = (1/K)\Sigma\, x_{k*}$ is subtracted from each row vector in the data matrix and the resulting centered matrix is denoted by $\mathbf{X}^{(0)}$. It may be considered a zeroth-order residual matrix, the mean vector constituting the zeroth-order model.

The PCA model is a *bilinear model* composed of two different sets of vectors that form an outer product:

$$\mathbf{M}^{(A)} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p}_i^T \tag{1}$$

The latent variables $\{\mathbf{p}_i\}$ (the principal components) are eigenvectors of the covariance matrix $\mathbf{X}^{(0)\dagger}\mathbf{X}^{(0)}$. $\mathbf{M}^{(A)}$ is an estimate of the centered data matrix, based on the A first principal components. The corresponding residual matrix is denoted by $\mathbf{X}^{(A)} = \mathbf{X}^{(0)} - \mathbf{M}^{(A)}$. The squared Frobenius norm of the Ath-order residual matrix is used to measure the error in the PCA model:

$$s_A = \sum_i \sum_j \left(X_{ij}^{(A)}\right)^2 \tag{2}$$

In agreement with the definition of variance in the exact data matrix, s_A is considered to be the variance left unexplained by the PCA model using A components. Thus, the percentage of explained variance attributed to component A is given by $100\%(s_A - s_{A-1})/s_0$.

For use in interpretation, it is important to be able to understand and classify the pattern of covariance which each principal component conveys. To this end, one may use simple two-dimensional plotting techniques. In the *loading* plot, the contribution from each of the original variables is plotted for a single pair of latent variables at a time. This is a graphical presentation of the linear expansion coefficients, called *loadings* in PCA nomenclature, for the two principal components. The loading plot visualizes the relationship between the different variables. Variables close to the center of the plot are said to be of zero loading and do not contribute to the construction of the principal component model. If a variable is substantially displaced from the origin only along the direction of component i, this means that the variable is mostly associated with PC i. Variables which are displaced in the same (opposite) direction along the direction of component i are positively (negatively) correlated along PC i. Variables close together in the loading plot have a strong covariance.

In the *score* plot, a pair of principal components, usually those which explain most of the variance, are used as axes, and the reaction pathway is projected onto the resulting two-dimensional plane. The score plot visualizes how the molecular properties at different geometries along the reaction

path are related to each other. Configurations with similar properties are close to each other. Similar to the loading plot, if a structure has a high score only along principal component i, this means that the properties at the given point along the reaction path is mostly explained in terms of PC i. Structures that are located in the center, that is, at zero score values, are well represented by the mean properties along the reaction path and do not contribute to the principal component model.

It can be instructive to superimpose a loading plot on top of the corresponding score plot, an arrangement frequently referred to as a *biplot*. The variables that are located close to a projected point along the reaction path are the most important for describing this part of the reaction. Biplots are not explicitly produced in the present paper, but the concepts involved in their use provide a convenient mental tool for the interpretation of PCA results.

SELECTED DESCRIPTORS

In this work, we have tried out four different descriptors of the structures at each computed intrinsic reaction coordinate (IRC) point. Two of these descriptors are geometrical parameters, whereas the remaining two descriptors quantify the electronic structure in terms of density matrix elements.

Geometrical Parameters

Two types of geometrical descriptors are used: (1) interatomic distances only and (2) zeta matrix coordinates: interatomic distances, angles, and dihedral angles.

For the first kind of parameter, at each geometry $\mathbf{G}^{(k)}$ along the IRC, a property row vector x_{k*} was constructed from the set of all $N(N-1)/2$ interatomic distances.

The second kind of coordinates are more complicated. First, two different types of variables occur in the same data set: distances and angles. Second, angles are cyclic variables, which cause problems for linear models. A cyclic space has the topology of a multidimensional torus, which is different from the flat Euclidean space which forms the basis for most multivariate methods. This has important implications for how the distance between points should be calculated. For instance, an angle of 5° is closer to 350° than it is to 180°. The problems can be avoided if the data at hand occupy a small part of the N-torus that can be approximated by a flat space or by embedding the cyclic space in a Euclidean space of larger dimension. The former approach is restricted to "well-behaved" systems, while the latter solution is less than ideal in that the number of variables is increased. However, based on the concept of embedding the unit circle in a two-dimensional plane, we have chosen to transform the angular variables into new, distance-like variables according to $x_i = \sin((\theta_i - \bar{\theta})/2)$. Each angle θ_i is considered to define a point $(\cos\theta_i, \sin\theta_i)$ on the unit circle, and the concept of an average point $(\bar{r}\cos\bar{\theta}, \bar{r}\sin\bar{\theta}) = (1/N)(\Sigma\cos\theta_i, \Sigma\sin\theta_i)$ lets us compute the average angle $\bar{\theta}$ from its inverse tangent. The transformed variable x_i is then seen to be the signed Euclidean distance between two points on the unit circle, namely, the points defined by the average angle and the angle θ_i. The sign is introduced in order to improve the resolution of the variations close to the mean angle. When used together with bond distances in a single principal component analysis, the transformed angles are further multiplied by a common factor (uniform scaling) to make the transformed angles vary on a similar scale as do bond distances. Based on data for the reactants, the scaling factor was determined by equating the range for scaled, transformed angles to the range of all changes in interatomic distances.

Properties Based on Mulliken Population Analysis

The distribution of the electronic density is described in terms of atomic charges and overlap populations as calculated by the Mulliken formalism, which is chosen because of its general availability and well-documented properties.

The atomic charges are frequently an order of magnitude larger than the overlap populations, and variations in the latter parameters would easily become masked if included in a single PCA model. Autoscaling, which forces each variable to display a variance of unity, would remedy the problem, but at the cost of exaggerating the importance of many variables originally of low variance. To avoid this, the atomic charges and the overlap populations are treated as two different descriptors and they are employed in separate principal component analyses.

For the Mulliken overlap populations, at each geometry $\mathbf{G}^{(k)}$ along the IRC, the property row vector x_{k*} was constructed from the set of all $N(N-1)/2$ interatomic overlap populations. However, overlap populations may be of any sign,

and this leads to difficulties in the analysis. For instance, a decreasing overlap population may be due to a reduced bonding if the overlap is positive or an increased antibonding interaction if the overlap is negative. To ensure that a decreasing value of a variable may be interpreted as less interaction, we chose to modify the overlap vectors by considering only the absolute values of overlap populations. Used in this manner, the absolute values of the overlap populations may be viewed as a transformation of the matrix of all interatomic distances, such that the importance of long and intermediate distances is greatly reduced. In addition to the influence of bond lengths, the overlap populations are of course also influenced by orbital changes such as hybridization and orientation.

For atomic charges as descriptor, the property vector simply consists of the N Mulliken atomic charges.

Presentation of the Chemical Reactions

Two chemical systems of differing complexity are chosen for investigating the prospects of the PCA method with respect to analyzing quantum chemical reaction data. First, multivariate analyses are performed on data for the propagation step in the industrially important Ziegler–Natta polymerization reaction, as calculated for an 18-atom model. Emphasis is put on demonstrating how the score and loading plot may be combined in order to extract chemical information. This particular reaction is well known from previous work and it is of particular interest to investigate the sensitivity of the multivariate approach in detecting the finer details. PCA is regularly used for classification in other areas of chemistry, and in the second example, two reaction pathways for the addition of hydrochloric acid to propene are compared. Here, we are exploring the potential of the multivariate approach for detecting similarities and differences between related reactions.

ZIEGLER–NATTA INSERTION REACTION

The Ziegler–Natta catalysts[18–20] consist of a main group metal alkyl or hydride, often designated as the cocatalyst, and a transition metal salt (e.g., $TiCl_4$). They are capable of polymerizing olefins into long molecular chains with both high speed[21–24] and stereospecificity.[25] The remarkable properties of these catalysts have inspired several theoretical studies; see, for example, refs. 26–29.

Here, the reaction mechanism proposed by Cossee[30] for the chain propagation step is adopted. According to this mechanism, a propagation step is initiated by the coordination of the monomer onto a vacant site on a five-coordinate, square-pyramidal metal complex, with at least one alkyl ligand, thus forming the so-called π-complex; see Figure 2 (top). The monomer subsequently inserts into the metal–alkyl bond via a four-center transition state involving the alkyl group, the alkene, and the transition metal.

The model catalyst chosen is the neutral, bimetallic compound $AlH_2(\mu\text{-}Cl)_2TiCl_2(CH_3)$, where the aluminum moiety models the cocatalyst. Including ethylene, this system contains 18 atoms and 130 electrons. The multivariate analyses

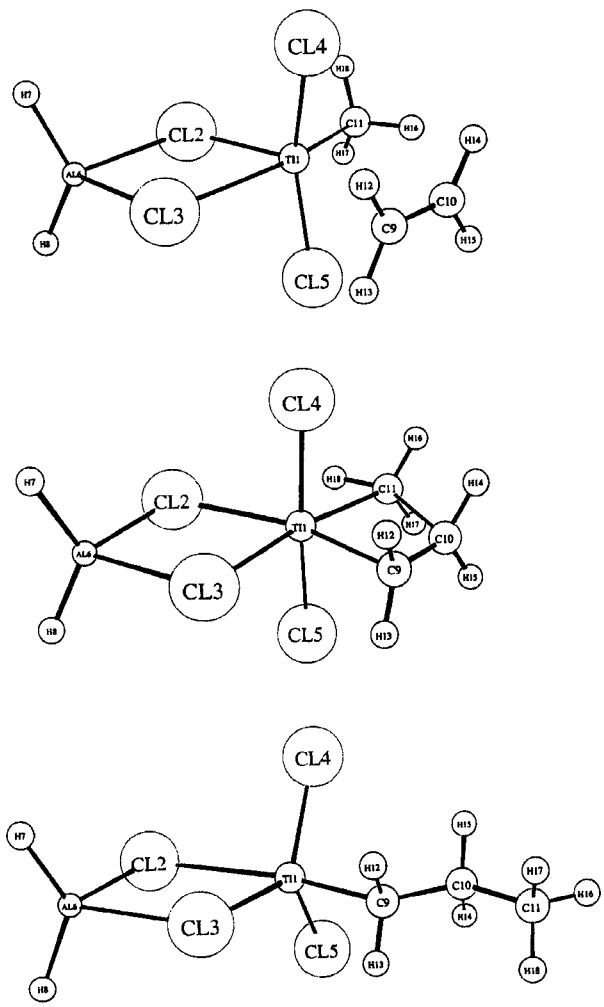

FIGURE 2. Ziegler–Natta system. Molecular complex (top), transition state (middle), and product (bottom).

are performed on the reaction data starting with the π-complex and leading to the final product.

ELECTROPHILIC ADDITION OF HCl TO PROPENE

Addition of haloacids to an unsymmetric alkene may give two products, depending on to which carbon the hydrogen is added. In general, these kinetically controlled additions obey Markovnikov's rule, which, in its simplest form, is that hydrogen adds to the ethylenic carbon which has more hydrogen atoms.[31-33] The rate-determining step is believed to be the formation of an intermediate carbocation, formed by protonation, and the usual rationalization of Markovnikovs rule is based on the stability of this intermediate.[34] The anti-Markovnikov reaction (AM) is then disfavored on account of the intermediate being a primary carbocation, whereas the Markovnikov (M) reaction path involves a more stable secondary carbocation.

The two reaction pathways start out from a common molecular complex, pictured in Figure 3 (top). From this point on, the reaction forks into the M and AM pathways. The transition state (TS) structures for the M (left) and AM (right) paths are shown in the middle of Figure 3. There corresponding products, 2-Cl-propane (left) and 1-Cl-propane (right) are shown at the bottom of Figure 3.

QUANTUM CHEMICAL DESCRIPTION OF THE REACTIONS

The self-consistent field (SCF) restricted Hartree–Fock (RHF) approximation was used in all quantum chemical calculations reported in the present paper.

Titanium was described by a valence triple-ζ basis set. The rest of the elements were described by valence double-ζ sets with an uncontracted polarization function (d) added for carbon and chlorine. In the calculations of electrophilic addition of HCl to propene, an uncontracted polarization function (p) was added to the hydrogen basis.

The reaction pathway calculations were carried out by first locating the transition states, and then integrating the intrinsic reaction coordinate (IRC) equation[35] in both forward and backward directions, using the Gonzales and Schlegel second-order method.[36,37] The step size used was 0.3 $(\text{amu})^{1/2}$ bohr. The GAMESS set of programs[1] was used in all the reported calculations, which were performed on the Intel Paragon XP/S at the National Massive Parallel Processing center, University of Bergen, and on the Intel Paragon A/4 at SINTEF, Trondheim.

Further details of the quantum chemical methods and basis sets may be found in ref. 29.

Analysis of the Ziegler–Natta Insertion Reaction

The reader is encouraged to consult Figure 2 for definitions of atomic labels. All analyses in this section are performed on data representing every fourth point along the computed reaction path.

GEOMETRY VARIABLES

An initial attempt was made to model the geometry changes by means of interatomic distances only. This was abandoned, mostly due to the inadequate representation of rotations. In order to reduce the number of variables and to improve the interpretability, we then chose to represent the molecular structures by a set of linearly independent internal coordinates, consisting of 17 bond lengths, 16 bond angles, and 15 dihedral angles. Prior to the PCA, all angles were transformed from cyclic to linear, distance-like variables, as detailed in Section 2.

Ninety-two percent of the variance in the internal coordinates is described by only two latent variables. The resulting score plot, Figure 4, reveals that the reaction path may be naturally divided into three distinct regions, labeled I–III. Furthermore, it is evident that the first principal component (PC1) mainly describes the difference between the reactant and the product, since they have almost the same score along the PC2 axis. This line of reasoning may also be used to conclude that PC2 is important for describing structures *intermediate* between reactant and product.

The chemistry taking place in regions I–III may be established by analyzing the composition of the most important principal components. The loading plot (not shown) turns out to be dominated by torsional angles describing rotations within the polymer chain. PC1 does, however, describe the formation of the expected new Ti-C and C-C bonds, as well as a rotation of the terminal ethyl group, on going from reactant to product. This rotation is seen to correlate with an interchange in the bridging bond distances [e.g. (CL2, Ti1) and (CL3, Ti1)].

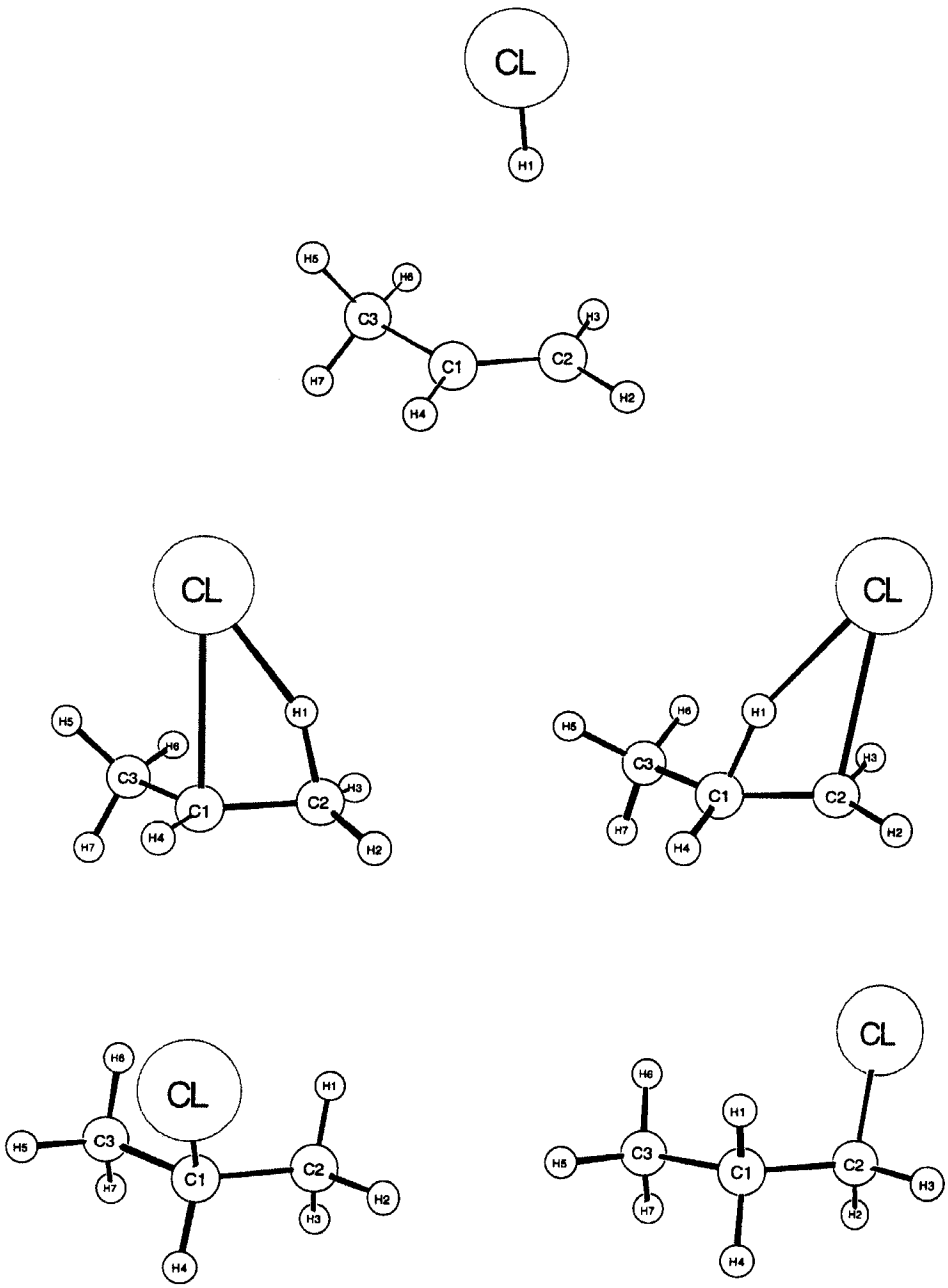

FIGURE 3. Electrophilic addition system. Molecular complex (top), M transition state (middle left), AM transition state (middle right), M product (bottom left), and AM product (bottom right).

The model thus provides indications of a "flipping" of the polymer chain as the reaction proceeds from a reactant complex with the methyl group positioned trans to CL3, to the final product with the propyl group positioned trans to CL2; compare Figure 2 (bottom).

Region II turns out to be dominated by the actual insertion reaction. By superimposing the score plot on the loading plot, it is evident that variables describing bond breaking and forming are located along the same direction in loading plot as the two-dimensional projection of the path in region II. Likewise, regions I and III are dominated by internal coordinates describing rotations of the end-methyl and end-ethyl in the polymer chain. However, several features of the reaction are very difficult to detect in the present PCA model. This includes the elongation of the eth-

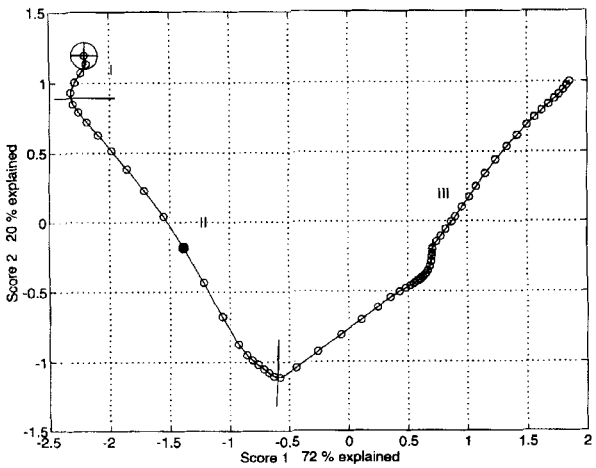

FIGURE 4. Score plot of the ZN insertion reaction. Internal coordinates are used as descriptors. ●, transition state; ⊕, molecular complex.

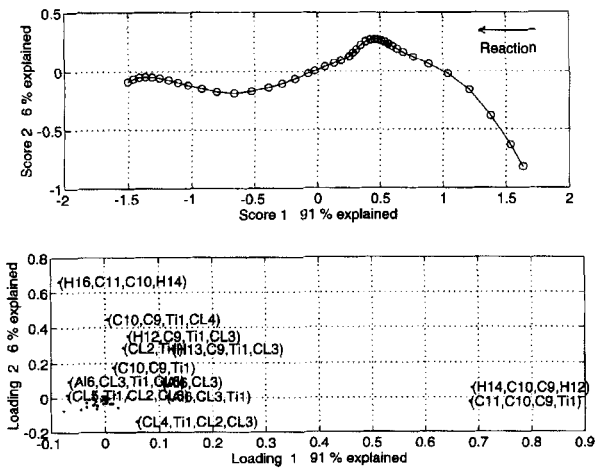

FIGURE 5. Score plot (top) and loading plot (bottom) of ZN insertion reaction with internal coordinates. Only the last part of the reaction, involving rotations of the propyl group, is included in this PCA model. ●, transition state; ⊕, molecular complex. In order to resolve overlapping labels in the loading plot, the corresponding data points are shifted slightly from their true positions. These shifts are small and do not affect the use of the figure.

ylenic C-C bond during insertion, as well as the α-agostic interaction known to take place in the transition region.[29] In view of the sectional character of the reaction path, as reflected in the score plot, it is likely that a more detailed picture of the reaction may be obtained by performing separate analyses of the different regions. The bond forming and breaking in region II is most likely to be adequately analyzed in terms of overlap populations, and this part is thus covered in the next section.

In the remaining part of the present section, region III is analyzed in terms of internal coordinates, which should be particularly suitable for rendering a detailed picture of the rotations. As pointed out, at the onset of region III, the insertion has already taken place. This is confirmed by the length of the ethylenic C-C bond, (C10, C9), which has increased to 1.55 Å, only 0.01 Å shorter than the newly formed C-C bond (C11, C10). The rest of the reaction is mainly dedicated to relaxation toward a trigonal bipyramidal complex, completing the catalytic cycle.

In the analysis of region III, two latent variables suffice to describe 97% of the variance; compare the upper part of Figure 5. It is evident from the loading plot in the lower part of this figure that PC1, explaining 91% of the variance, is dominated by the two torsions (H14, C10, C9, H12) and (C11, C10, C9, Ti1) which describe rotation of the end-ethyl. The largest contribution to PC2 is from torsion (H16, C11, C10, H14), which describes rotation of the end-methyl.

In region III, the reaction starts off at a PC1 score of 1.6. The curve changes rapidly in both PC1 and PC2 directions, implying that both components are important. Two concerted rotations take place: one about C11-C10 and the other about the C10-C9 bond. The methyl rotation about C11-C10 is eased as the ethyl rotation about C10-C9 is performed in the opposite direction. A staggered methyl conformation can thus be reached at an early stage. From a starting value in region III of 0°, (H16, C11, C10, H14) already reaches 57° four points later; compare the score plot. Next, the curve levels off and reaches a maximum in PC2 score. This maximum corresponds to a local minimum on the potential energy surface (PES) and is caused by a staggered conformation of the two methylene groups. At this point, the torsional angles describing the ethyl rotation, (C11, C10, C9, Ti1) and (H14, C10, C9, H12), are reduced by close to 60°. The low-energy conformation is reached sooner via a second rotation in the opposite direction of the ethyl rotation. Now it is the α-methylene group which rotates about C9-Ti1, as evidenced by the contribution to PC2 from variables (C10, C9, Ti1, CL4), (H12, C9, Ti1, CL3), and (H13, C9, Ti1, CL3); see Figure 5 (lower part). The latter two torsion angles increase by 13° in the first part of region III.

Past the local energy minimum, the importance of PC2 is reduced and much of the structural change can be described by the rotation about C10-C9, that is, by PC1. The total rotation about C10-C9 amounts to 180°, which is appreciated by comparing the middle and bottom structures in Figure 2—transition state of insertion and product, respectively. The reduction in PC2 score after the local minimum on the PES is mostly caused by a small back-rotation about C9-Ti1 to reach the C_s-symmetric product.

Finally, it is worth noticing that the fine synchronization and concert of some of the rotations in region III were *not* recognized in a movie[38] of the reaction until the PCA had drawn attention to these details.

MULLIKEN ATOMIC OVERLAP POPULATIONS

The upper part of Figure 6 shows the score plot based on absolute values of overlap populations for the reaction, all three regions included. PC1 describes the overall difference in overlap populations between reactant (the π-complex) and product, and explains 92% of the variance. Consequently, PC2 describes chemical interactions that are particular to the transition region. These interactions may be looked upon as a retardation of the reaction in the transition region and are seen to reach maximum importance shortly after the transition state.

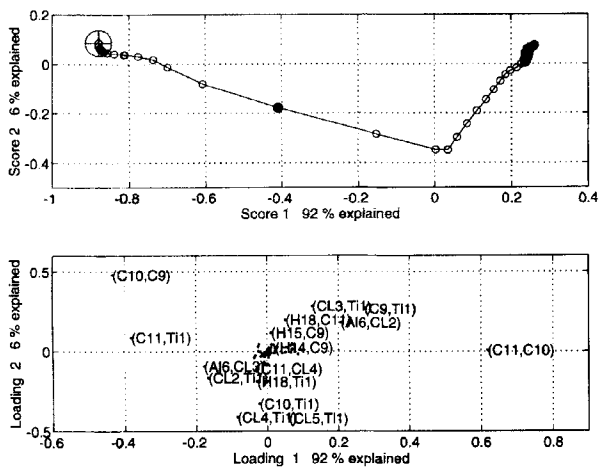

FIGURE 6. Score plot (top) and loading plot (bottom) of the ZN insertion reaction with Mulliken overlap populations as the structure descriptor. ●, transition state; ⊕, molecular complex. In order to resolve overlapping labels in the loading plot, the corresponding data points are shifted slightly from their true positions. These shifts are small and do not affect the use of the figure.

According to the loading plot (Figure 6, lower part), the most important contribution to PC1 is the overlap population (C11, C10) of the carbon–carbon bond formed during the reaction. This variable is strongly negatively correlated to variables describing bonds to be broken during the reaction, that is, the titanium–methyl bond (C11, Ti1) and the ethylenic π-bond (C10, C9). The variable corresponding to the new titanium–carbon bond (C9, Ti1) is found at a positive PC1 score of 0.3. Significant contributions from the titanium–chlorine–aluminum bridges are also detected. The populations of the two shorter bonds in the reactant, (CL2, Ti1) and (Al6, CL3), are decreasing and accordingly located at negative PC1 values. (CL3, Ti1) and (Al6, CL2) are located at approximately the same positions along the positive PC1 axis. The interchange in bond distances for the bridges, from reactant to product, is thus clearly reflected in the analysis of the overlap populations.

Three overlap populations are particularly important for describing the transition structures, namely, the two bonds to terminal chlorines, (CL4, Ti1) and (CL5, Ti1), as well as (C10, Ti1). These variables, located at large negative PC2 values, are increasing from reactant until two points past the transition state (cf. score plot), while the ethylenic (C10, C9) is found at the opposite extreme of PC2. The maximum values of the overlap populations (CL4, Ti1) and (CL5, Ti1) coincide with maximum bond lengths for these bonds. Thus, the large overlaps in the transition region are probably caused by a combination of rehybridization and increased electron density on the metal. The increase in (C10, Ti1), on the other hand, follows the reduction of the corresponding interatomic distance. PC2 is also seen to describe an increase in (H18, Ti1) in the first part of the reaction. (H18, Ti1) is negatively correlated with (H18, C11), which reflects an α-agostic interaction in the transition region.

The bonds in the two titanium–chlorine–aluminum bridges are also negatively correlated in PC2. During the first part of the reaction, PC2 and PC1 describes opposite tendencies for the bridges. However, the score plot shows that PC1 has the larger impact until a zero PC1 score is reached, ensuring the interchange of the bridges. At positive PC1 scores, the interchange is described by both components.

According to PC2, variables (H18, Ti1) and (C10, Ti1) covary and increase in the first part of

the reaction. However, the corresponding overlap populations are negative, implying that the overlaps get increasingly negative during the reaction. The negative populations may very well be caused by artifacts from the Mulliken population analysis. This seems at least to be the case for (H18, Ti1), as a reduced population (H18, C11) confirms the agostic and, thus, bonding interaction between titanium and H18.

PCA based on the inverse of the 19 most important interatomic distances, including (H18, Ti1) and (C10, Ti1), turned out to give almost the same information about the reaction as the one presented here based on overlap populations. The most significant difference was that the bonds to terminal chlorines, (Cl4, Ti1) and (CL5, Ti1), turn out to be *negatively* correlated with the agostic (H18, Ti1). This is expected as the bonds to the terminal chlorines reach maximum lengths in the transition region.

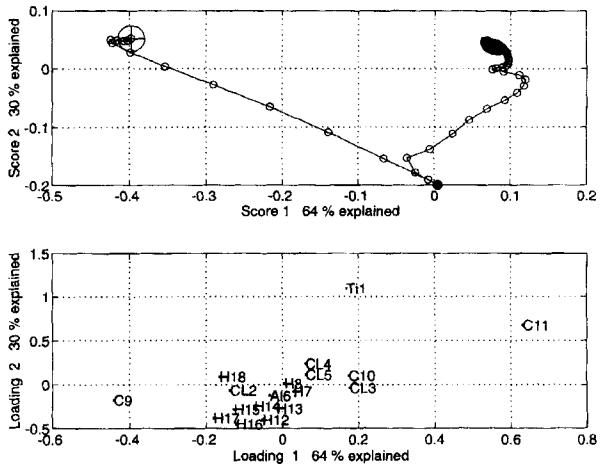

FIGURE 7. Score plot (top) and loading plot (bottom) of the ZN insertion reaction with Mulliken charges as the structure descriptor. ●, transition state; ⊕, molecular complex.

MULLIKEN CHARGES

In most of the analyses of the Ziegler–Natta reaction, the first principal component describes the overall reaction; see, for example, the score plot based on overlap populations, Figure 6. However, this pattern is not confirmed in the PCA model based on atomic charges. In order to simplify both the interpretation and the presentation of the charge fluctuations, it was decided to perform an orthogonal rotation of PC1 and PC2, so as to obtain a single component describing the difference between reactant and product. It should be noted that the rotation does not change the two-component PCA model; it only affects the partitioning into two components. Hence, the total explained variance is conserved. However, the rotated components are no longer *principal* components since they do not correspond to directions of maximum variance in the data, and they will be referred to as latent variables. The rotated score and loading plots are shown in Figure 7, and it can readily be verified that latent variable 1 (LV1) does describe the overall reaction. Furthermore, LV2 is important for describing the charge distribution in the transition region.

Turning to the loading plot (Fig. 7, bottom), LV1 is seen to be dominated by a negative correlation of the charges of C9 and C11. C11 starts out as a methyl carbon (charge $-0.77e$) directly attached to titanium and ends up as a terminal methyl carbon (charge $-0.48e$) in a propyl chain. In the molecular complex, C9 starts out as an ethylenic carbon (charge $-0.38e$) and becomes a methylene carbon (charge $-0.59e$) directly attached to titanium. The other carbon initially participating in the ethylenic double bond, C10, is transformed into a methylene carbon in the propyl moiety in the product. It suffers a smaller change in atomic charge, going from $-0.41e$ to $-0.33e$, and this is reflected in the lower contribution to LV1. However, in the transition region, C10 experiences some charge fluctuations that are not modeled by any of the two latent variables. In the previous section on overlap populations, some emphasis is put on the interchange in the bond lengths of the two chlorine bridges connecting titanium and aluminum. This process may be seen in LV1, in terms of a negative correlation between the charges of the two bridging chlorines, CL2 and CL3.

LV2 consists mainly of a negative correlation between Ti1 and C11 on one hand, and most of the hydrogen atoms on the other hand. The deviation of H18 from the other hydrogen atoms warrants further checking of other variables related to this atom. Indeed, overlap populations clearly show that H18 is subject to an agostic interaction in the transition region, as detailed in a previous section.

Most of the projection of the path between the molecular complex and the transition state resembles a straight line. Only the first small part, which is a simple rotation with only small effect on the charge distribution, deviates from this trend. The variance in the charges in the first part of the insertion reaction (prior to the transition state) can

be attributed to a reduction in the charges on C9 and Ti1, described by LV1 and LV2, respectively. The increased electron density on Ti1 in the transition region is mainly at the expense of the hydrogen atoms, but some electron density is also withdrawn from Al6 and C9. The charge on C11 increases only slowly prior to the transition state, since the increase predicted by LV1 is offset by a similar decrease predicted by LV2. A similar but smaller effect is seen for C9, as LV2 dampens the reduction in charge in the transition region. Past the transition state, the contribution from LV2 is much larger and the charge on titanium increases to become slightly higher than in the reactant. Most of the increase in C11 also takes place in the second half of the reaction.

MULLIKEN CHARGES AND OVERLAP POPULATIONS COMBINED

In a previous section it was made clear that the actual insertion reaction takes place in region II (cf. Figure 4), and data from this section were made subject to a detailed analysis using both Mulliken charges and absolute values of overlap populations as descriptors. However, the combined analysis gave little in excess of the separate analyses presented above. For example, PC1 is very similar to the sum of the first principal components from each of the separate analyses, shown in the lower part of Figures 6 and 7. The reduction in charge of carbon C9 in the double bond thus occurs concerted with the weakening of the π-bond and the Ti–methyl bond. At the same time, the charge on the methyl carbon (C11) increases with the formation of a new single C-C bond. The second principal component is dominated by the charge on titanium. It is possible to discern most of the other contributions to PC2 found in the separate analyses, except for some of the details.

Comparison of Two Related Reaction Pathways

In this section, we report our experience with principal component analysis as a tool for comparing two closely related reactions: the Markovnikov (M) and anti-Markovnikov (AM) addition of hydrochloric acid to propene. Several ways of performing such a comparison were tried out, using Mulliken atomic charges as descriptors. Based on these results, a single strategy was adopted and applied to several other kinds of descriptors. The reader may want to consult movies of the M and AM reaction paths.[38]

MULLIKEN CHARGES

Initially, separate principal component analyses were performed for the Markovnikov and anti-Markovnikov reaction paths. The atoms were labeled according to their position in the reactant molecular complex; refer to Figure 3 (top) for definitions. For both reactions, the first principal component is dominated by differences in atomic charges between the transition state (TS) on one hand and the stable reactants and product on the other; see Figure 8. The second principal component distinguishes between the molecular complex and the product. While this approach is useful for obtaining information about the charge flow in each reaction, it is less useful for *comparing* the two reactions. The reason is that the score plots cannot be directly compared, since they refer to different sets of axes.

One way to get around this problem is to perform the analysis on the combined sets of data for both reactions. In this case, a single set of principal components are obtained and the two reactions may be plotted in a common score plot; see Figure 9. The Markovnikov reaction is well described by the first principal component. The second PC describes most of the differences between the M and AM paths, and explains 33% of the total variance. However, in the loading plot (not shown), it ap-

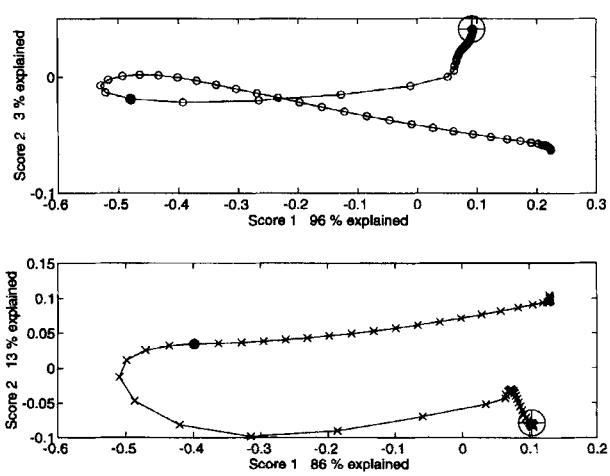

FIGURE 8. PCA score plots for the Markovnikov (top) and anti-Markovnikov (bottom) reactions analyzed separately. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reactant-consistent labeling.

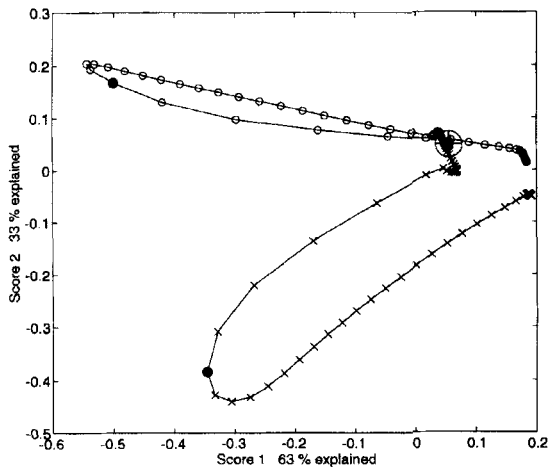

FIGURE 9. PCA score plots for both the M and AM reactions modeled together. Atomic charges as descriptor. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reactant-consistent labeling.

pears that the main role of PC2 is to point out pairs of atoms having the same functionality in the two reactions. The ethylenic carbons, C1 and C2, make an obvious example, in that C1 eventually binds to the chlorine in the M reaction, whereas C2 has this functionality in the AM reaction. Since this information is available *prior* to the analysis in our case, it can be taken into account by relabeling the atoms with emphasis on functionality in the reaction and such that it makes sense chemically to compare properties of atoms carrying the same label. We will denote this choice of atomic labels for *reaction-consistent* labels, in contrast to the *reactant-consistent* labels originally chosen.

The switch to reaction-consistent labeling reveals a complication of general interest: a single molecular configuration, in our case the reactant molecular complex, may correspond to two different property vectors, that is, two different points in the score plot. The difference between the two points represents in some sense a nil vector, in that it amounts to a zero difference in chemical properties. This obscures the analysis of the score plot, since it is difficult to distinguish chemically important displacements from those with no chemical significance. One suggestion for amending the situation is to remove any contribution from the nil vector by projection in the data matrix. However, this will remove a lot of valuable information as well. In fact, for the present system, such a procedure will erase any information about the difference between charges for the two ethylenic carbons. A more useful procedure is to subtract the nil vector from each property vector for one of the reactions, forcing the representations for the common molecular complex to coincide.

This latter strategy is adopted in the remaining comparisons of the two pathways. The labels are chosen to mirror the functionality of each atom in the M reaction, making C1 the carbon which eventually binds to chlorine and C2 the carbon which receives an additional hydrogen. The corresponding labeling for the AM reaction is shown in Table I. For the hydrogens in propene, it is not possible, or even interesting, to assign functionalities on an atom-by-atom basis. Rather, average "atoms" are introduced, such that H4 (HHm, MeH) represents the average properties of the hydrogens bonded to C1 (C2, C3). HHd represents the differences in properties between the two hydrogen atoms bonded to C1 (C2) in the M (AM) reaction.

Turning to the score plot (see the upper part of Figure 10), it is evident that the roles of PC1 and PC2 are similar to what they were before relabeling the atoms. However, the importance of PC2 has been reduced to 9% explained variance, implying that some 25% of the variance before relabeling was due to the choice of labels. It is likely that the information left in PC2 is to a large extent of chemical importance.

The first principal component describes most of the charge flow along the Markovnikov reaction pathway. From the molecular complex, the reaction proceeds in the negative PC1 direction (cf. both score and loading plots, Figure 10). C1, located at a large positive PC1 value in the loading plot, is negatively correlated with chlorine. Thus, the charge on chlorine gets increasingly negative, whereas the opposite trend holds true for C1 and

TABLE I.
Definitions of the Reactant- and Reaction-Consistent Labels Used in the Electrophilic Addition System [HHm = ½(H2 + H3), HHd = H2 − H3, MeH = ⅓(H5 + H6 + H7)].

Reactant-Consistent Labels	Reaction-Consistent Labels	
	M	AM
H1	H1	H1
CL	CL	CL
C1	C1	C2
C2	C2	C1
H2, H3	HHm	H4
H2, H3	HHd	HHd
H4	H4	HHm
C3	C3	C3
H5, H6, H7	MeH	MeH

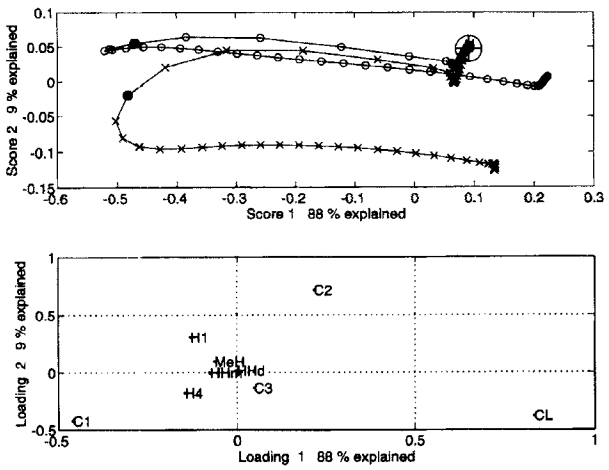

FIGURE 10. PCA score plot (top) and loading plot (bottom) of the M and AM reactions analyzed together, using atomic charges as descriptor. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reaction-consistent labeling. In order to resolve overlapping labels in the loading plot, the corresponding data points are shifted slightly from their true positions. These shifts are small and do not affect the use of the figure.

the hydrogen(s) attached to it (H4). The hydrogen from HCl (H1) gets increasingly positive and stabilizes a higher electron density on C2. Past the transition state, charge flows in opposite directions as compared to before TS, eventually leading to an atomic charge on chlorine which is somewhat higher than in the molecular complex. The AM reaction is to a large extent described by the same model, but there is a significant difference in the transition region. Here, the reaction proceeds to a large extent in the negative PC2 direction (cf. the score plot). According to the loading plot, the most important contribution to PC2 is a negative correlation between C1 and the chlorine, located at large negative PC2 values, and C2. This implies that electrons are withdrawn from chlorine and C1, and seemingly transferred to the other ethylenic carbon (C2). As the projection of the last part of the AM path can be observed at low PC2 scores, it is clear that this difference persists all the way from transition states to products. This implies that there are features apart from the chlorine which governs the charge distribution for the ethylenic carbons. While the Markovnikov product has a much higher (i.e., less negative) charge on the carbon bonded to chlorine (C1) than the other ethylenic carbon (C2), the importance of PC2 for the anti-Markovnikov reaction suggests that the difference is much smaller in the AM product. The immediate explanation for this is the role played by the methyl group. In 2-Cl-propane, C1 is bonded to two relatively electron-rich methyl groups, whereas in 1-Cl-propane, C2 is the central carbon atom.

GEOMETRICAL ANALYSIS

The analysis is based on interatomic distances for all computed points along the pathways for both the Markovnikov and the anti-Markovnikov reactions, and reaction-consistent labeling was prepared as previously described. Three principal components are extracted to account for close to 100% of the variations in interatomic distances during the reactions. In Figure 11, the reaction pathways are plotted with respect to these three PCs, and the corresponding two-dimensional score plots are shown as projections. The projection onto the PC1-PC3 plane almost coincides for the two reactions, whereas the remaining score plots (PC1-PC2 and PC2-PC3) show substantial differences between the two reactions. These simple observations allow us to focus on PC2 for describing most of the differences between the Markovnikov and anti-Markovnikov reactions.

The first principal component (PC1) describes 83% of the variance in the distances, and from the loading plot in Figure 12 it is evident that PC1 describes the highly correlated approach of H1 and CL to carbon C2 and C1, respectively, with attached hydrogen atoms HHm and H4. The bond in hydrochloric acid, represented by (CL, H1) at a

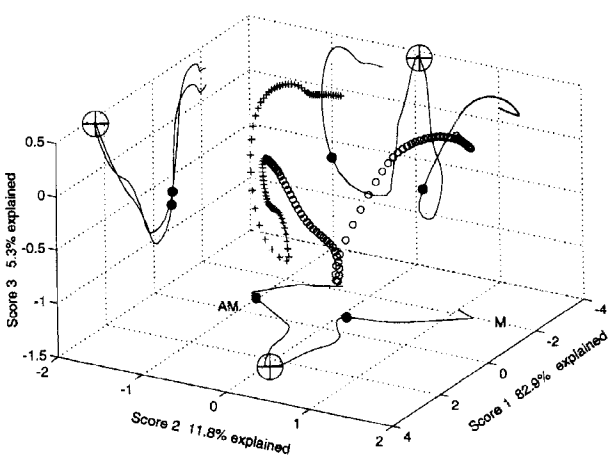

FIGURE 11. PCA scores 1, 2, and 3 of the M and AM reactions analyzed together. Interatomic distances as structure descriptors. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reaction-consistent labeling.

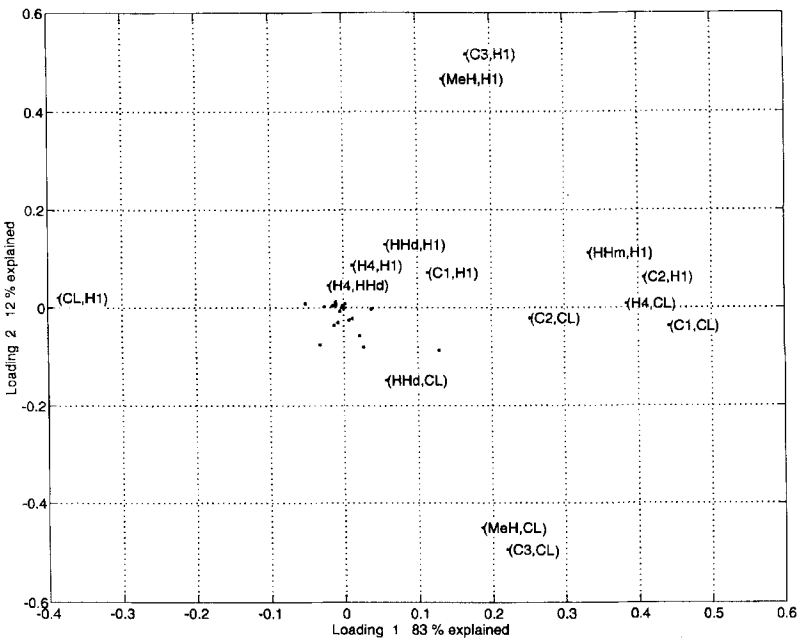

FIGURE 12. Loading plot (1 vs. 2) for PCA on both the M and AM reactions together with distance matrices as the structure descriptors. Reaction-consistent labeling. In order to resolve overlapping labels in the loading plot, the corresponding data points are shifted slightly from their true positions. These shifts are small and do not affect the use of the figure.

negative PC1 value, is stretched in concert with the approach to propene. PC1 describes, to a large extent, the progress of the reactions and is close to linearly related to the reaction coordinate lengths s; see Figure 13. This finding suggests that the scores for PC1 may be used as a common "pseudo" reaction path length for the two reactions, allowing

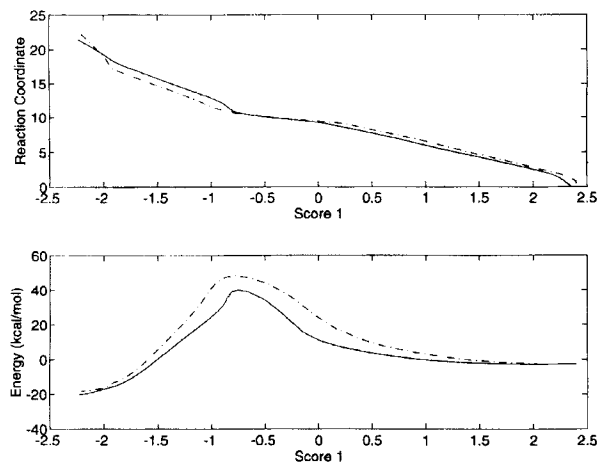

FIGURE 13. Score 1 versus the reaction coordinate for the M (−) and AM (− ·) reactions (top). Score 1 versus energy (kcal / mol) for the M (−) and AM (− ·) reactions (bottom).

their energy profiles to be compared in a single plot; see lower half of Figure 13. Note that a plot of energy versus reaction path length s would have been questionable, since the two reactions are governed by different reaction coordinates.

The second principal component (PC2) explains 12% of the total variance and describes, as noted above, the differences between the M and AM paths. This interpretation of PC2 is confirmed by its composition, as pictured in the loading plot (Figure 12). PC2 conveys a single piece of information, namely, the differential approach to the methyl moiety made by the two atoms from HCl. This is evident because the distances between H and methyl and CL and methyl, located at a PC1 value of ~ 0.2, are negatively correlated. Hence, while PC2 differentiate between the two atoms in HCl, PC1 describes the general approach of both H1 and CL to the methyl. The interpretation of PC2 will be used as we turn to the reaction paths as plotted with respect to PC1 and PC2; see Figure 14.

In the region marked by I in the score plot, the M and AM scores are almost mirror images of one another (about the PC1 direction), and the approach to the methyl group is in accordance with the expectations for each reaction. However, in

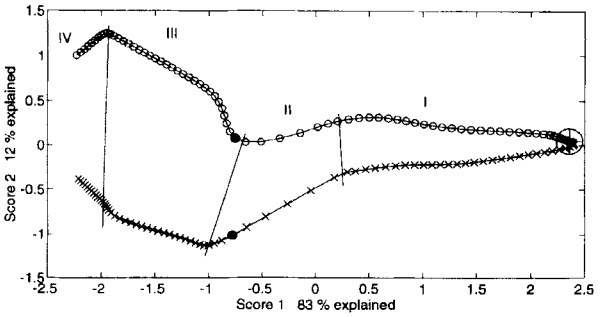

FIGURE 14. Score plot (1 vs. 2) for PCA on both the M and AM reactions together with distance matrices as the structure descriptors. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reaction-consistent labeling.

regions II and III, very similar translations occur in both reactions. In region II, prior to the transition states, the H1 hydrogen is approaching the methyl group. Then, in region III, just after the TS, it is the chlorine which moves toward the methyl group in both reactions. This clearly shows that, even though the hydrogen H1 and chlorine CL eventually end up on opposite carbons in the two reactions, a very similar mechanism is active in the central part of the reactions. In both cases, initially it is the hydrogen which plays the active role and attacks the propene. Only as the transition states are passed, the chlorine moves on toward its end destination. Thus, from the PC1-PC2 score plot, important indications of an electrophilic mechanism are available. In region IV, the two reactions again display opposite behavior with respect to the second principal component. Clearly, this part describes rehybridization of the carbon with a methyl attached. The change to sp^3 hybridization makes the methyl group bend away from the new atom introduced, which is chlorine in the case of the Markovnikov path.

No distances within the propene molecule contribute to any of the two most important latent variables. This not only shows that the distortions of propene are rather small, but that the elongation of the ethylenic bond in propene is camouflaged by the translations. In order to also model changes in the distances at the order of 0.05 Å, a third variable is extracted. It explains 5% of the total variance and mainly models deviations in the relative orientation of the two reactants from the linear-transit-like approach which is described by PC1. In a plot of the reaction pathways versus PC2 and PC3 (Figure 11), it becomes clear that it is mainly structures in the vicinity of the transition states which have such deviations. The corresponding loading plot (not shown) reveals that it is the details in the positioning of H1 that are important in this respect. The contributions from the elongation of the ethylenic C-C bond are still barely discernible.

A comparison of the geometrical changes during several related reactions may conceivably be performed in terms of zeta matrix variables, that is, both bond angles, dihedral angles, and bond distances. However, it is less straightforward to identify internal coordinates that are comparable on account of their similar functionality during the reaction. On this account, such an analysis was not performed.

MULLIKEN OVERLAP POPULATIONS

A very high fraction (0.98) of the variations in overlap populations is explained by a single latent variable. From the loading plot in Figure 15, it is seen that all major chemical effects, such as bond formation (C2, H1), (C1, CL) and bond breaking (C2, C1), (CL, H1), are described by this component. Second, from the score plot (Figure 16) it is evident that PC1 describes the difference between reactants and product. Third, similarly to what was found for interatomic distances, the contribution from PC1 increases linearly with the path coordinate. One may therefore conclude that 98% of the variations in overlap populations may be accounted for simply in terms of linear interpolation between reactants and product.

The second component accounts for most of the difference between the TS and the equilibrium states, but represents only 2% of the variance. The PC1-PC2 loading plot (not shown) reveals a negative correlation between the overlap populations (C1, CL) and (C2, CL). The effect is to retard the bonding between chlorine and carbon 1 somewhat in the transition region compared to the linear-transit model advocated by PC1. Further, PC2 adds some bonding between CL and C2, making the interaction between chlorine and the ethylenic carbons less asymmetric. Also, the breaking of the ethylenic double bond is somewhat delayed compared to an interpolating model.

Neither of the two most important principal components distinguish well between the M and AM pathways, as judged from the score plot (Figure 16). The third PC, explaining less than 1%,

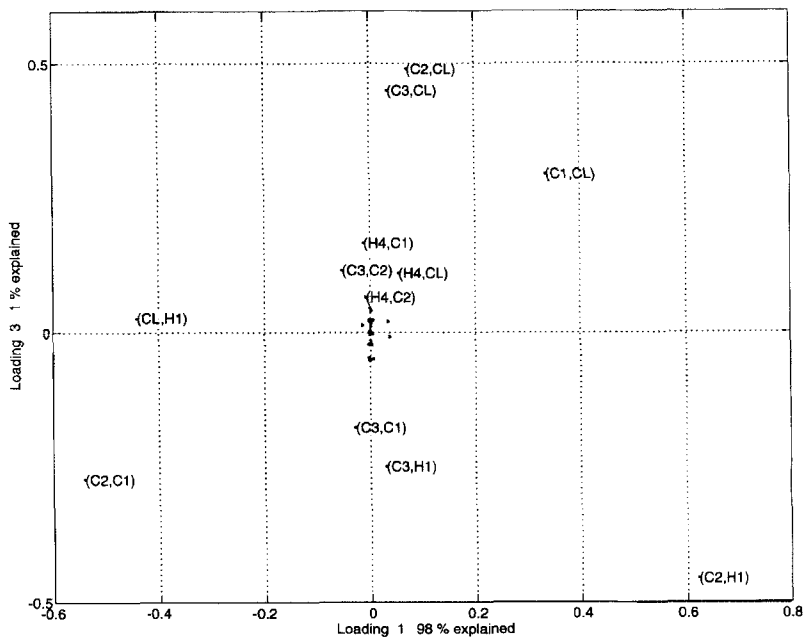

FIGURE 15. Loading plot (1 vs. 3) for PCA on both the M and AM reactions together with overlap populations as the structure descriptors. Reactant-consistent labeling. In order to resolve overlapping labels in the loading plot, the corresponding data points are shifted slightly from their true positions. These shifts are small and do not affect the use of the figure.

appears to describe differences related to rehybridization in the final stage in the reactions. This is confirmed in the loading plot (Figure 15), which points to an increased nonbonding repulsion between the methyl group and chlorine for the Markovnikov reaction and a corresponding change involving H1 for the AM reaction.

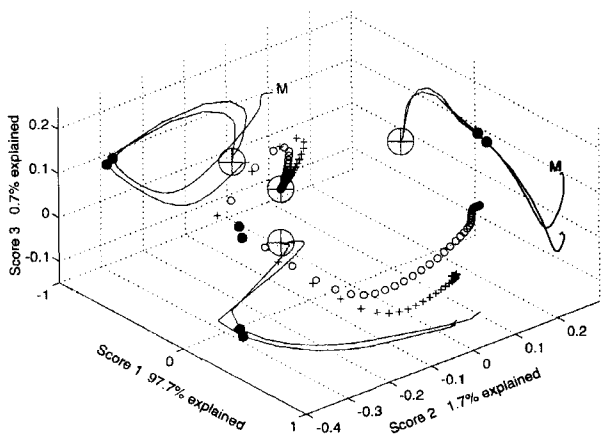

FIGURE 16. The three-dimensional plot of scores 1, 2, and 3 for PCA on both the M and AM reactions together with overlap populations as the structure descriptors. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reactant-consistent labeling.

COMBINED CHARGES AND INTERATOMIC DISTANCES

In order to follow the covariance of charge flow with geometrical changes, a principal component analysis was performed on a data set consisting of all interatomic distances and Mulliken charges for both reaction pathways. The charges were scaled uniformly, as described for angular variables in Geometrical Parameters, to ensure that they vary on the same scale as do changes in distances. The resulting score plot is shown in Figure 17 and indicates that the third component, explaining 8% of the variance, contains information about the differences between the two reaction pathways. From the loading plots, there seems to be little covariance between the two kinds of descriptors: charges and distances. Rather, the two main components may be viewed as originating from the rigid rotation of the two PC1 components from the separate analyses, fixed at right angles to one another. Similarly, PC3 is composed of the PC2s of the separate analyses discussed above. However, there is one instance of charge–distance covariance

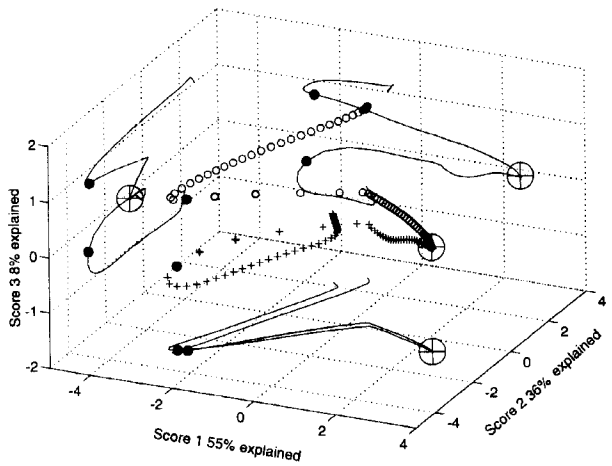

FIGURE 17. The three-dimensional plot of scores 1, 2, and 3 for PCA on both the M and AM reactions together with distance matrices and atomic charges as the structure descriptors. ○, M; +, AM; ●, transition state; ⊕, molecular complex. Reaction-consistent labeling.

which persists in both the PC2-PC3 and PC1-PC3 (Figure 18) loading plots, and that is made by the C2 charge and the distance between the methyl moiety and the added hydrogen (H1). This signifies that the buildup of negative charge on C2 is proportional to the proximity of hydrogen H1 to the methyl moiety. It is likely that the role of the H1–methyl distance is to flag which of C1 and C2 is bonded to the methyl carbon (C3). This lends support to our interpretation of the analysis based on Mulliken charges only, namely, that bonding to the methyl group is of decisive importance for the charge distribution between the two ethylenic carbon atoms.

Discussion

The main features of a reaction are accessible through score and loading plots based on any of the properties considered in this work. However, the ease with which a given feature is detected and interpreted may be very dependent on the kind of property modeled. For instance, interatomic distances are useful for describing translations, whereas rotations are most easily detected in terms of internal coordinates. Overlap populations are well suited to describe bond formation and bond breaking and, to some extent, also local translations.

Interatomic distances may covary even though the atoms in question are widely separated and do not participate in a direct chemical interaction.

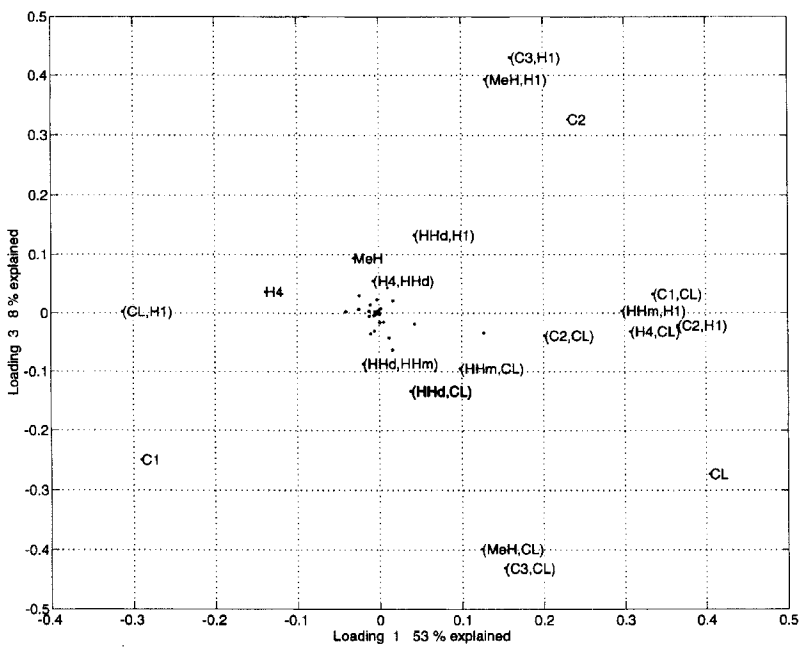

FIGURE 18. Loading plot (1 vs. 3) for PCA on both the M and AM reactions together with distance matrices and atomic charges as the structure descriptors. Reaction-consistent labeling. In order to resolve overlapping labels in the loading plot, the corresponding data points are shifted slightly from their true positions. These shifts are small and do not affect the use of the figure.

This is a drawback in that it may lead to covariance patterns of little interest, but is helpful in terms of seeing the synchronized restructuring of a large system. The same significance is attached to a relative translation of 0.2 Å whether the fragments in question are at a distance of 10 or 1 Å. This implies that if large translations occur, then they will quite effectively mask any local distance changes, at least in the first few principal components. In order to avoid this, one may simply remove the large translations from the analysis or apply a distance-dependent renormalization of the distances. The latter idea is attractive in that one may design the renormalization function as to open a window to a specified range of distances. Our experience is that it is difficult to find a compromise between simplicity in the analysis and conservation of chemical information.

It is, in some sense, the overlaps of valence basis functions which define what is the chemically interesting distance. Thus, one suggestion may be to select a normalized, valence atomic orbital of spherical symmetry for each atom and use the overlap integrals between these orbitals as a distance measure. An even more attractive alternative, employed in our analyses, is to model the Mulliken overlap populations. Apart from the distance between two atoms, there are, of course, also chemical effects influencing a given overlap population. This proves useful, in that a single set of properties contains both geometrical and chemical information.

Internal coordinates represent the geometrical structure in terms of parameters that are easily interpreted. The interpretability may be further enhanced by using an oversaturated coordinate set, that is, more than $3N - 6$ variables, such that every degree of freedom of chemical interest is represented by a well-designed coordinate. An advantage of both interatomic distances and Mulliken overlap populations is that they are easily and uniquely defined once the geometry and basis sets are given. This is not the case for general internal coordinates. Even if a zeta matrix is available from the geometry optimizations, it may be less than ideal for modeling purposes. It is important that the internal coordinates varies in a smooth manner with the geometry changes. This is a matter of special concern in connection with dihedral angles in close to planar atomic arrangements.

There exist other tools for analyzing geometrical changes during a reaction, and a very useful tool is to make a movie of the reaction. This is more laborious than performing PCA, but it gives a unique insight. However, for large reacting systems, it may be difficult to sort things out in a movie. Also, small but synchronized geometry changes may elude attention. Thus, a useful approach would be to combine a movie with principal component analyses.

The charge flow in a chemical system is well represented by atomic charges and overlap populations. This has made Mulliken population analysis a standard tool for quantum chemists. For the present systems one would probably obtain an adequate overview of the atomic charges from univariate plots. However, the usefulness of multivariate methods increases with the size of the system and the complexity of the properties to be modeled. For instance, the structure of the overlap populations is difficult to analyze by inspection even for modestly sized systems.

A less attractive parametrization of the total or partial electron density is a grid representation. The reason is that the amount of data for a moderately sized reactive system will outgrow any fast storage device even if data compression is applied. Bader[39] has developed a compact representation of the electron density in terms of so-called critical points. This concept is compatible with his theory of atoms in molecules, that is, atomic identities are preserved, and seems to be promising in connection with principal component analysis.

The molecular orbitals themselves exemplify a very complex property which would be of interest to include in a principal component analysis. One may consider limiting the analysis to only certain sets of molecular orbitals, such as the set of all occupied valence orbitals. Special care must be taken to ensure that the orbitals remain comparable during the reaction. This precludes the use of canonical orbitals, since two orbitals may swap character at avoided level crossings or at points of high symmetry. The preferred approach may be to use localized orbitals and will be the focus in future work.

The question of sensitivity of multivariate methods in the present context is an important one. Having observed the major features of the reaction, one goes on looking for more subtle details. These may be expected to show up in the second or third component. However, the trust one may have to tendencies in the higher components is related to the percentage explained variance. At some level, the principal components start to model noise. It is our experience that the loading plots of the higher components should be used as a torch,

directing attention toward a small subset of variables. The tendencies indicated in the score plot should always be checked for in the original data to ensure that the linear multivariate model is capable of representing the features in question. Actually, it may be that the model is not very good, but it may still be useful in that attention is drawn to a variable with strong and nonlinear variations.

The crux of comparing several reactions by means of multivariate techniques is the choice of units that are comparable between the reactions. In this work, it is chosen to focus on similar functionality, as discussed in detail in Mulliken Charges. Usually, such a mapping between several reactions will be incomplete in that a given functionality is present only in some of the reactions. It is, of course, of particular importance also to include these groups in the analysis.

The problem with obtaining comparable variables makes internal coordinates less useful. It is likely that specific rotations and out-of-plane bends may be represented by the corresponding angles, but the zeta matrix approach already presents insurmountable difficulties for comparing two simple reactions.

Types of multivariate methods other than PCA may be of use for quantum chemical data sets. One powerful method is the partial-least-squares (PLS) regression,[6-8] which creates a regression model between the independent and dependent data, based on latent variables. New score and two types of loading vectors are created that correspond to directions in the data set that maximize the covariance between the independent and dependent variables. This was tried out on distance data for the Markovnikov addition reaction, with the energy as the dependent variable. It was found that the principal components from PCA are close to optimal for predicting the energy of the system and, accordingly, PLS gave results very similar to PCA. The reason for this may be that the structures constitute a reaction coordinate and, as such, are selected because of their ability to contribute to steep ascent/descent in energy.

Common for PCA and PLS is that the points along the reaction coordinate are considered to constitute a *set*, rather than a *sequence*. A suggestion for exploiting the latter property is to perform PCA on data from a few neighboring IRC points at a time and then slide this "window" along the reaction path. The resulting score and loading plots should be combined to score and loading movies, offering a unique resolution of both geometrical and electronic structure changes.

Conclusions

Principal component analysis has been applied to *ab initio* computed data for chemical reaction pathways. The results are encouraging. However, it may be necessary to combine several kinds of structural descriptors in order to get a complete picture of a reaction. Furthermore, it has proven fruitful to model on subsections of the reaction path in order to increase the resolution of the multivariate model. Multivariate methods should thus become a useful tool in the analysis of calculated reaction path data for medium to large reacting systems.

Acknowledgments

Financial support from The Norwegian Academy of Science and Letters and Den Norske Stats Oljeselskap a.s. (VISTA, grant V6415) is gratefully acknowledged, as is a grant of computing time from the Norwegian Supercomputing Committee (TRU). Professor Olav M. Kvalheim and Professor Leif J. Sæthre are thanked for their continued interest in this work. We are indebted to M.Sc. Bjarne G. Herland and Dr. Jeffrey J. Gosper for visualization of the reaction paths.

References

1. M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, *J. Comput. Chem.*, **14**, 1347 (1993).
2. M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, C. Y. Peng, P. Y. Ayala, W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople, *Gaussian 94*, Gaussian, Inc., Pittsburgh, PA, 1995.
3. E. J. Baerends, D. E. Ellis, and P. Ros, *Chem. Phys.*, **2**, 41 (1973).
4. H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, New York, 1989.
5. S. Wold, K. Esbensen, and P. Geladi, *Chemom. Intell. Lab. Syst.*, **2**, 37–52 (1987).
6. A. Höskuldsson, *J. Chemom.*, **2**, 211–228 (1988).

7. S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, *SIAM J. Sci. Statist. Comp.*, **5**, 735–743 (1984).
8. A. Lorber, L. Wangen, and B. R. Kowalski, *J. Chemom.*, **1**, 19–31 (1987).
9. F. P. S. C. Gil, A. M. Amorim da Costa, R. E. Bruns, and J. J. C. Teixeira-Dias, *J. Phys. Chem.*, **99**, 634 (1995).
10. E. Suto, H. P. Martins F°, and R. E. Bruns, *J. Mol. Struct. (THEOCHEM)*, **282**, 81 (1993).
11. E. Suto, M. N. Ramos, and R. E. Bruns, *J. Phys. Chem.*, **97**, 6161 (1993).
12. T. Horiuchi and N. Gō, *Proteins*, **10**, 106 (1991).
13. A. Kitao, F. Hirata, and N. Gō, *Chem. Phys.*, **158**, 447 (1991).
14. R. D. Cramer, III, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.*, **110**, 5959–5967 (1988).
15. B. K. Alsberg, *Chemom. Intell. Lab. Syst.*, **8**, 173–181 (1990).
16. M. Stone and P. Jonathan, *J. Chemom.*, **7**, 455–475 (1993).
17. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
18. K. Ziegler, Belgian patent 533362, Nov. 1953.
19. G. Natta, *J. Polym. Sci.*, **16**, 143 (1955).
20. G. Natta, P. Pino, and G. Mazzanti, *Gazz. Chim. Ital.*, **87**, 528 (1957).
21. W. Kaminsky, K. Külper, and S. Niedobla, *Makromol. Chem. Macromol. Symp.*, **377**, 3 (1986). A turnover rate of 200,000 s^{-1} is reported.
22. K. Soga, M. Ohgizawa, and T. Shiono, *Makromol. Chem. Rapid Commun.*, **10**, 503 (1989).
23. N. Kashiwa and J. Yoshitake, *Makromol. Chem. Rapid Commun.*, **3**, 211 (1982).
24. T. Keii, M. Terano, K. Kimura, and K. Ishii, *Makromol. Chem. Rapid Commun.*, **8**, 587 (1987).
25. P. Pino and R. Mülhaupt, In *Transition Metal Catalyzed Polymerizations*, R. P. Quirk, Ed., Cambridge University Press, Cambridge, MA, 1983, vol. 4, p. 23.
26. H. Kawamura-Kuribayashi, N. Koga, and K. Morokuma, *J. Am. Chem. Soc.*, **114**, 2359 (1992).
27. H. Weiss, M. Ehrig, and R. Ahlrichs, *J. Am. Chem. Soc.*, **116**, 4919 (1994).
28. R. J. Meier, G. H. J. van Doremaele, S. Iarlori, and F. Buda, *J. Am. Chem. Soc.*, **116**, 7274 (1994).
29. V. R. Jensen, K. J. Børve, and M. Ystenes, *J. Am. Chem. Soc.*, **117**, 4109 (1995).
30. P. Cossee, *J. Catal.*, **3**, 80 (1964).
31. V. Markovnikov, *Liebigs Ann. Chem.*, **153**, 228 (1870).
32. G. Jones, *J. Chem. Ed.*, **38**, 297 (1961).
33. N. Isenberg and M. Grdinic, *J. Chem. Ed.*, **46**, 601 (1969).
34. S. H. Pine, *Organic Chemistry*, 5 ed., McGraw-Hill, New York, 1987, p. 375.
35. K. Fukui, *Acc. Chem. Res.*, **14**, 363 (1981).
36. C. Gonzales and H. B. Schlegel, *J. Phys. Chem.*, **94**, 5523 (1990).
37. C. Gonzales and H. B. Schlegel, *J. Chem. Phys.*, **95**, 5853 (1991).
38. Movies are available on the World Wide Web site: http://www.uib.no/kj/theo/anim/anim.html
39. R. F. W. Bader, *Atoms in Molecules, A Quantum Theory, International Series of Monographs on Chemistry*, vol. 22, Oxford University Press, Oxford, 1994.